Rejuvenating the Italian WordNet: upgrading, standardising, extending

Antonio Toral Stefania Bracale Monica Monachini Claudia Soria
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Pisa, Italy
{antonio.toral, stefania.bracale, monica.monachini, claudia.soria}@ilc.cnr.it

Abstract

This paper reports on recent activities carried out within the KYOTO project aimed at enhancing the Italian WordNet Language Resource. On the one hand we study the formalisation of this lexicon according to the LMF ISO standard and explore its application into a real-world scenario by means of representing it in the WN-LMF dialect. On the other hand, we report on a semiautomatic procedure to upgrade the connections of the lexicon to WordNet, which obtains over 98% accuracy.

1 Introduction

The goal of the KYOTO project[1] (Vossen et al., 2008) is the construction of a system for facilitating the exchange of information across cultures, domains and languages. This system is expected to allow people in communities to define the meaning of their words and terms in a shared Wiki platform so that it becomes anchored across languages and cultures but also so that a computer can use this knowledge to detect knowledge and facts in text. Whereas the current Wikipedia uses free text to share knowledge, KYOTO will represent this knowledge so that a computer can understand it. The system is being developed for the domain of environment. For example, the notion of environmental footprint will become defined in the same way in all these languages but also in such a way that the computer knows what information is necessary to calculate a footprint. With these definitions it will be possible to find information on footprints in documents, websites and reports so that users can directly ask the computer for actual information in their environment.

This endeavour presupposes the sharing of lexical databases and knowledge bases, both general and domain-related, under the form of lexical repositories and ontologies. The lexical resources that will be integrated in KYOTO are wordnets for the English, Dutch, Italian, Basque, Spanish, Chinese and Japanese languages. Special-domain wordnets and ontology will be developed: they are to be seen as a plugin extension of the generic wordnet and ontology. These extensions contribute to the development of the Global Wordnet Grid[2], which is an initiative to anchor many wordnets for different languages and cultures to a shared ontology backbone.

As in KYOTO the integration of resources is viewed as a need, the use of formats that facilitates interoperability is essential. Interoperability allows an easier integration among general domain lexicons sharing the same structure (i.e other wordnets) and domain lexicons, but, more importantly, eases the integration of resources with different theoretical and implementation approaches, such as the ones being used within the project: Web 2.0 sources (DbPedia), species taxonomies (Species2000) and ontologies (DOLCE, SUMO, SIMPLE). There is no means to speak about interoperability if not paired with standards: they are bound to be the communicative channel by means of which diverse data, resources, formats, and models can interact on a common ground, in a controlled way.

This paper reports on recent activities aimed at enhancing the Italian WordNet (IWN) (Alonge et al., 1999), according to the needs posed by the KYOTO project. On the one hand we study the formalisation of this lexicon according to a standard and explore its application into a real-world scenario by means of tailoring the standard to the practical requirements. The adoption of a standard will allow IWN to communicate with the other resources available in the KYOTO architecture. On the other, we upgrade the connections of IWN to the Inter-lingual Index (ILI) (Vossen, 1998) to the

[1] http://www.kyoto-project.eu

[2] http://www.globalwordnet.org

latest version of the English WordNet (Fellbaum, 1998). This will allow a better interaction of IWN with the rest of wordnets of the project because it will be able to get corresponding senses by means of two different versions of the ILI.

The rest of the paper is organised as follows. Next section discusses the standardisation process followed to convert IWN to the LMF standard and to its dialect WN-LMF. After that, we report on the upgrade of IWN's connections to the ILI from WN 1.5 to the last version available, 3.0. Finally, we draw some general conclusions.

2 Standardisation

2.1 LMF

The Lexical Markup Framework (LMF) (Francopoulo et al., 2008) (ISO 24613, 2008) is an ISO standard for the representation of LRs. The goals of LMF are to provide a common model for the creation and use of LRs, to manage the exchange of data between and among them, and to enable the merging of a large number of individual resources to form extensive global electronic resources.

LMF has been chosen as representation format because it gathers experiences and harmonization efforts started by the interested community in the '90s. This format for lexical resource representation has now reached a high level of sophistication, theoretical consensus, and official international standard status, being ratified as an ISO standard (ISO 24613, 2008). LMF was specifically designed to accommodate as many models of lexical representations as possible. Purposefully, it is designed as a meta-model, i.e. a high-level specification for lexical resources defining the structural constraints of a lexicon. It is organised around two main components:

- The core package, i.e. a structural skeleton to represent the basic hierarchy of information in a lexicon, under the form of core classes of objects and relations.

- A set of modular extensions to the core package, i.e. additional classes and relations required for the description of specific types of lexical resources. Available extensions include morphology, syntax, semantics, multilingual notations, paradigm classes, multiword expression patterns and constraint expressions. Mutual dependencies among the various extensions hold.

Before being issued as an official ISO standard, LMF has passed a range of officially needed stages and has been extensively discussed and commented in a wide community comprising both academia and industry. LMF is thus mature enough to be taken as "the" choice when coming to selecting a standardised format for the representation and encoding of computational lexicons. Time is ripe now to start assessing LMF, providing the community with real examples of use.

A procedural routine has been developed in order to convert from the IWN specific XML format to LMF. The main difference found between both formats is that while in the specific one the information regarding the sense, synset and ILI relations are hold from a common ancestor ("WORD_MEANING"), in LMF they belong to different elements.

Let us present a sample from the specific IWN format:

```
<WORD_MEANING ID="AG#44455" PART_OF_SPEECH="AG">
 <GLOSS>che si può abbassare</GLOSS>
 <VARIANTS>
  <LITERAL LEMMA="abbassabile" SENSE="1"/>
 </VARIANTS>
 <INTERNAL_LINKS>
  <RELATION TYPE="liable_to" ID="75" INV_ID="75">
   <TARGET_WM ID="34802" PART_OF_SPEECH="V"/>
  </RELATION>
 </INTERNAL_LINKS>
 <EQ_LINKS>
  <RELATION TYPE="eq_synonym" ID="1" INV_ID="1">
   <TARGET_WM ID="r#345085"/>
  </RELATION>
 </EQ_LINKS>
</WORD_MEANING>
```

It follows the corresponding LMF code, separated in three blocks (lemma and sense, synset and ILI):

```
<LexicalEntry id="LE_abbassabile_a">
 <Lemma>
  <feat att="partOfSpeech" val="a"/>
  <feat att="writtenForm" val=
   "abbassabile"/>
 </Lemma>
 <Sense id="abbassabile_1"
  synset="ita-15-44455-a"/>
</LexicalEntry>

<Synset id="ita-15-44455-a">
 <Definition>
  <feat att="gloss" val="che si può
   abbassare"/>
 </Definition>
 <SynsetRelation targets="ita-15-34802-v">
  <feat att="relType" val="liable_to"/>
 </SynsetRelation>
</Synset>

<SenseAxis id="sa_0" synsets=
 "ita-15-44455-a eng-15-345085-r">
 <feat att="relType" val="eq_synonym"/>
</SenseAxis>
```

2.2 WN-LMF

Wordnet-LMF (WN-LMF) is an LMF dialect tailored to encoding of lexical resources adhering to the WordNet model of lexical knowledge representation. No real attempt has been made so far in order to fully apply LMF to wordnet-like lexicons: WN-LMF is an example of the practical use of LMF in a real-world application (Soria et al., 2009). The KYOTO project represents an ideal test case for this format: going beyond the level of toy examples it allows to make a crash test, as the various resources need to be fully integrated. This will put us in the position to both have a preview on any problems we might encounter and make LMF standard easy to adopt. More importantly, we will be able to convince people that there is a good reason to convert their legacy formats, by showing its usefulness and efficiency.

WN-LMF fully complies with the standard LMF as for its general framework. It builds on the representational devices made available by LMF and tailors them to the specific content requirements of the WordNet model of lexical knowledge representation. LMF library provides the hierarchy of lexical objects with structural relations among them. The Data Category library provides the elementary descriptors to be used in combination with the structural elements, necessary to represent lexical information (Francopoulo et al., 2006). Figure 1 shows a general diagram of WN-LMF.

2.2.1 WN-LMF overall design

The main conceptual components of WordNet-like lexicons that need to be represented in LMF are the following:

- Synsets, variants and synset relations, including information about synset identifiers and sense-keys;
- Domain attribution, linking to ontologies, administrative information;
- Interlingual information, i.e. mapping of synsets in a given language to Interlingual Index (ILI).

The LMF semantic package naturally lends itself to the representation of wordnet-like resources, since it already contains lexical objects devised for the representation of synsets, their associated gloss and examples, variants, and synset relations.

Expression of WordNet-related types of information (such as synset relations, external sources linked to wordnets) falls into the realm of LMF Data Categories, which are by definition either selectable from the pre-defined standard registry or custom-defined. The WN-LMF format, accordingly, has defined a Data Category Selection, necessary to fully represent the various wordnets to be integrated in KYOTO. Examples of custom Data Categories are values for describing synset relations, inter-lingual relations, for identifying external resources and their associated nodes, etc. For the sake of better parsing efficiency, in WN-LMF, Data Categories are represented by means of XML attributes and values instead of nested lexical objects. As an example consider the following sample of LMF code:

```
<Lemma>
  <feat att="partOfSpeech" val="n"/>
  <feat att="writtenForm" val="abbadia"/>
</Lemma>
```

and its equivalent in WN-LMF:

```
<Lemma partOfSpeech="n"
  writtenForm="abbadia"/>
```

By explicitly naming the attributes, we also make a stronger claim about the features and properties of the structure of a wordnet. This will enforce better compatibility and interoperability across the many wordnets for different languages that are available. In this respect, the WN-LMF DTD implementation has to be seen as a dialectal variant of the LMF DTD. Motivation behind this choice is to reach efficiency, while keeping adherence to standards.

2.2.2 The WN-LMF core component

The WN-LMF core package component provides the structural skeleton to represent the basic hierarchies of the lexicon.

KYOTO WordNets are represented as a grid of lexicons: *LexicalResource* is the container for all of them. A specific set of lexical objects is devoted to record general information about the lexical resource.

The lexical resource, besides the monolingual lexicons, contains the interlingual correspondences which are grouped in a section *SenseAxes* which is separated from the lexicons proper and contains inter-lexicon correspondences only.

Lexicon contains a monolingual resource, instantiated as a set of *LexicalEntry* instances. This element is a container for representing a lexeme in a lexicon. A *LexicalEntry* element contains the

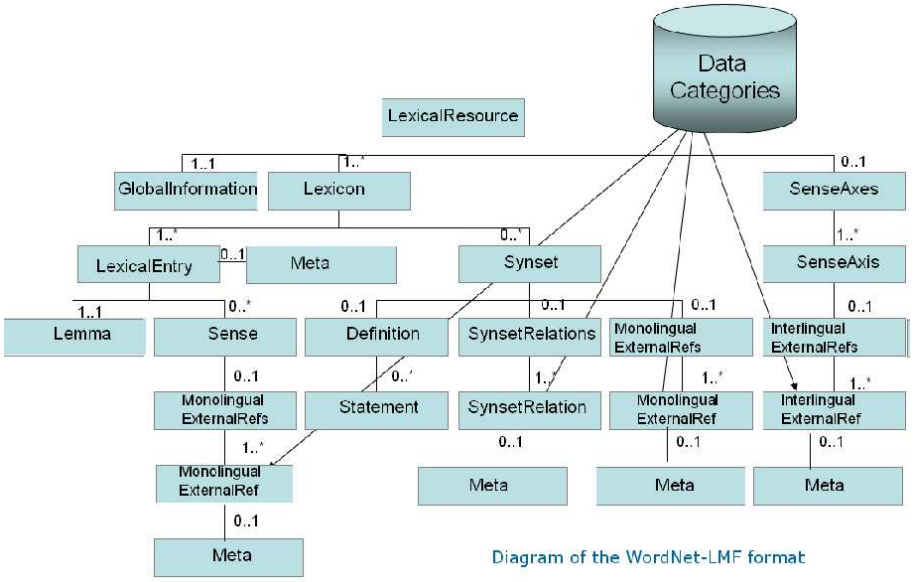

Figure 1: WNLMF diagram

basic building blocks: lemma and senses. *Lemma* represents a word form chosen by convention to designate the lexical entry, whereas *Sense* represents one meaning of a lexical entry. For wordnet representation, this triplet is used to represent the variant(s), or literal(s) of a synset.

MonolingualExternalRef represents linking between a *Sense* or *Synset* and another resource, be it a knowledge organisation system, a database, or another lexical resource. Mapping among different versions of the same resource, reference to external information, such as mapping onto entries of another lexical database and or referencing additional sources can be dealt with by the *MonolingualExternalRef* object.

When linked to a *Sense* element, it can be used to express mapping between the sense and its correspondent in another lexical resource (such as in the Dutch Cornetto database). In the particular case of the representation of English Princeton WordNet, *MonolingualExternalRef* serves as a representational device to express the Sense Key. When linked to the *Synset* element, then *MonolingualExternalRef* allows to encode reference to the domain and/or one or more links to an ontological system.

2.2.3 The WN-LMF semantic component

The Semantic component is in charge of describing information about a wordnet synset by means of the *Synset* element. A *Synset* clusters senses of different *LexicalEntry* instances within the same part of speech. The element *Definition* allows to represent the gloss associated with each synset. Relations between synsets are codified by means of *SynsetRelation* elements (represented by means of XML attributes), one per relation.

A set of harmonized KYOTO Data Categories has been defined. This is to be used in conjunction with the *SynsetRelation* elements for representing the various relations holding between synset. This Data Category library, for the sake of coherence, is being maintained as a centralized repository. This option has been followed in order to enforce better compatibility and interoperability across the many monolingual wordnets.

MonolingualExternalRef, which is used to represent linking between the lexical resource and another resource, when linked to the *Synset* element, allows to encode reference to the domain and/or one or more links to an ontological system.

2.2.4 The WN-LMF multilingual component

The Multilingual notation component is used in KYOTO for expressing interlingual correspondences. This component is designed as an independent package in order not to overload the representation of monolingual lexicons. The model is based on the notion of “Axes” that link synsets pertaining to different languages. For the purposes of creating a grid of WordNets linked via Interlingual Index, the *SenseAxis* device is specifically suited to implement approaches based on an interlingual pivot. Any *SenseAxis* element groups to-

gether monolingual synsets that correspond one to another by means of a particular type of relation.

The *SenseAxis* element is a means for grouping together synsets belonging to different monolingual wordnets that correspond one to another and share the same equivalence relation (e.g. a synonymy or near_synonymy relation) to a pivot synset, which by convention is an English one. This is a compact way of encoding correspondences among wordnets, avoiding to have several LanguageX-English single correspondences.

InterlingualExternalRef is used in WN-LMF to express a linking between a *SenseAxis* instance and an external system such as an ontology, and represents the means to anchor a multilingual group of synsets to an ontological node. Its intended use, thus, is to provide a representational device to link a group of synsets from different wordnets to the same ontological concept.

3 Upgrading

IWN was originally linked to version 1.5 of the ILI. In this section we report on a semiautomatic procedure carried out in order to update these links to the last version of WN at the time being, 3.0.

We take advantage of the automatic mapping sets between pairs of WN versions[3](Daudé et al., 2000). These mapping sets connect every combination of WN version pairs in both directions. E.g. for the version pair 1.5. and 3.0. there are two mapping sets, one from 1.5. to 3.0. and another one from 3.0. to 1.5. For each synset in the source version, the mapping sets provide the equivalent synset(s) in the target version together with a confidence score. Each mapping follows the following format:

```
synset_source [synset_target weight]+
```

An example taken from the WN 1.5 to 3.0 mappings is:

```
2728-n 4258-n 0.222 4475-n 0.778
```

which means that the synset "2728-n" of WN 1.5 is mapped to two synsets of WN 3.0, "4258-n" with confidence 22.2% and to "4475-n" with confidence 77.8%.

From the two directional mapping sets for our version pair (1.5. and 3.0.) we have created a bidirectional mapping set which follows the following format:

```
synset15 synset30 weight15->30 weight30->15
```

If a mapping is not present in one of the two directions we mark its weight as -1. These are the mappings for the source synset "2728-n":

```
2728-n       4258-n     0.222    1
2728-n       4475-n     0.778    1
2728-n    5217061-n    -1       1
```

An advantadge of using this bidirectional mapping over using a directional one can be seen in this example. If the directional mapping would be used to upgrade the ILI connections, for the synset "2728-n" there are two target candidates, whilst taking into consideration the bidirectional mapping, a third additional candidate is found.

When using these mappings to upgrade the links to ILI three cases can arise:

- There is a one to one equivalence. We select a subset randomly and check it manually to calculate the accuracy of the automatic mappings.
- There is no equivalence. We analyse why no equivalence is found and create a connection manually.
- There is a one to n equivalence (where n>1). These mappings need to be manually disambiguated.

IWN contains 50,308 synsets. From these, 106 are not connected to ILI while the rest are mapped to a total of 57,164 ILI synsets. From these, table 1 shows the number of synsets that fall into each of the aforementioned cases when using different mappings schemes. These are a directional (Dir) scheme and two bidirectional, one following an union (Bidu) and the other following an intersection pattern (Bidi).

ILI synsets	Dir	Bidu	Bidi
Total	57,164		
No equivalence	1,897	1,897	2,021
1-to-1 equivalence	54,817	42,614	53,133
1-to-1 equiv. (dir)	-	-	1,800
1-to-n equivalence	450	12,653	210

Table 1: Distribution of synsets with different mapping schemes

The next subsections report in more detail for the different cases. We have chosen the *Bidi* mapping scheme because it is the one that requires us to disambiguate less 1-to-n equivalences.

[3] http://www.lsi.upc.es/~nlp/tools/mapping.html

3.1 One-to-one equivalences

We have randomly selected a subset of 100 mappings of this type for each Part-of-Speech, i.e. adjectives (a), adverbs (r), nouns (n) and verbs (v). These mappings have been manually checked in order to evaluate the accuracy of the automatic mapping procedure. Results are shown in tables 2 and 3. The "total" scores in both tables normalise the score obtained for each Part-of-Speech by the number of occurrences for each PoS, see equation 3.1.

$$\frac{\sum_{pos\in(a,r,n,v)} num_{pos} * acc_{pos}}{num_a + num_r + num_n + num_v} \quad (1)$$

Part-of-Speech	Accuracy
Adjective	96%
Adverb	98%
Noun	99%
Verb	99%
Total	98.77%

Table 2: Results of 1-to-1 bidirectinal mappings

Part-of-Speech	Accuracy
Adjective	97%
Adverb	100%
Noun	99%
Verb	99%
Total	98.68%

Table 3: Results of 1-to-1 directinal mappings

The accuracy obtained for the 1-to-1 mappings is therefore very high, above 98% in average for both types of mappings. The performance for adjectives is slightly lower than for the others Part-of-Speech.

Errors occur seldom and regard very fine grained distinctions. Consider the example of WN 1.5 synset "35605-a" (quiet) which has not gloss but is connected through a "similar to" relation to synset "35448-a" (dormant, inactive) with gloss "of e.g. volcanos; temporarily inactive". The synset is mapped to WN 3.0 synset "43615-a" (quiet) with gloss "of the sun characterized by a low level of surface phenomena like sunspots", instead, the correct mapping would be "40909-a" (quiescent) with gloss "being quiet or still or inactive".

3.2 Ambiguous and empty equivalences

Both the ambiguous and empty equivalences have been manually resolved. Regarding the disambiguation task, we have applied the following disambiguation pattern: for each ambiguous concept we have selected the most appropriate one. This choice has been carried out in different steps. If the meaning of the term was unknown then we have looked it up in the IWN web interface[4]. Using the MCR interface[5], we have looked for the WN 1.5 and the WN 3.0 corresponding synsets. The most similar WN 3.0 synset has been selected by consulting different types of information related to each synset such as, its variants, its hyperonymy chain, etc.

The empty equivalences have been resolved with a different methodology: in the first step, for each empty entry, we have found its English correspondent by using various English dictionaries. In a second step, we have searched the WN 3.0 synsets that contain as a variant the translation obtained. If this entry has been found in WN 3.0 and it corresponded to the same Italian semantic concept expressed in IWN, then the code of this synset has been linked. If the meaning corresponded exactly, the type of relation chosen was EQ_SYNONYM, while if the meaning was similar but presented slight differences, then the type of relation chosen was EQ_NEAR_SYNONYM. Otherwise, if no unique correspondence has been found, then no connection has been created. We note that the most complex disambiguation task concerns some Part-of-Speech entries, such as adjectives and adverbs. There are adjectives in IWN that only exist as nouns in English (e.g. accusato/accused), and some adjectives that in English are only found as verbs (past participle) e.g. illustrato/illustrated. There are also cases of adverbs for which no correspondence was found in English.

4 Conclusions

This paper has reported on two recent activities that regard the extension, standardisation and upgrade of IWN.

With respect to the standardisation, we have studied and developed the conversion of this lexicon into the LMF ISO format. Furthermore, we

[4] http://wordnet.ilc.cnr.it/

[5] http://www.lsi.upc.es/~nlp/meaning/demo/demo.html

have discussed the implications of using the resulting resource in a real-world NLP scenario. We have devised the creation of a LMF dialect, WN-LMF, in order to increase efficiency while keeping adherence to the standard.

On the other hand, we have carried out an upgrade of the ILI links of IWN. We have followed a semiautomatic approach that takes advantage of existing automatic mappings between pairs of WN versions and checks manually only those mappings which are ambiguous or whose confidence scores are low. A contribution of this paper is an empirical evaluation of the automatic mappings, which obtain accuracy values higher than 98%.

An indirect yet useful contribution is the availability of manually disambiguated mappings between WN1.5 and WN3.0[6]. These could be exploited by WNs for other languages that are linked to WN1.5 (e.g. those developed in the framework of EuroWordNet) in order to upgrade their connections.

Acknowledgements

This work has been partially funded by the EU Commission under the project KYOTO (ICT-2007-211423). We thank German Rigau for his valuable advice regarding the automatic WordNet mappings.

References

Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, and Adriana Roventini. 1999. The Italian Wordnet, EuroWordNet Deliverable D032D033 part B5. Technical report.

Jordi Daudé, Lluís Padró, and German Rigau. 2000. Mapping wordnets using structural information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*, Hong Kong.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Gil Francopoulo, Monica Monachini, Thierry Declerck, and Laurent Romary. 2006. The relevance of standards for research infrastructure. In *LREC 2006, Workshop Towards Research Infrastructures for Language Resources*. European Language Resources Association (ELRA).

[6]Freely available at `http://www.dlsi.ua.es/~atoral/#Resources`

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2008. (forthcoming) Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation Journal*.

ISO 24613. 2008. Languages Resources Management – Lexical Markup Framework (LMF), rev.15 ISOTC37SC4 FDIS. [Online; accessed 25-March-2008].

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-lmf: fleshing out a standardized format for wordnet interoperability. In *IWIC '09: Proceeding of the 2009 international workshop on Intercultural collaboration*, pages 139–146, New York, NY, USA. ACM.

Piek Vossen, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Federico Neri, Remo Raffaelli, German Rigau, Maurizio Tesconi, and Joop VanGent. 2008. Kyoto: a system for mining, structuring and distributing knowledge across languages and cultures. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Piek Vossen. 1998. Eurowordnet a multilingual database with lexical semantic networks.